Neurocomputing 184 (2016) 232–242

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Auto-encoder based dimensionality reduction

Yasi Wang, Hongxun Yao*, Sicheng Zhao

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
3 August 2015
Accepted 4 August 2015
Available online 27 November 2015

Keywords:
Auto-encoder
Dimensionality reduction
Visualization
Intrinsic dimensionality
Dimensionality-accuracy

ABSTRACT

Auto-encoder—a tricky three-layered neural network, known as auto-association before, constructs the "building block" of deep learning, which has been demonstrated to achieve good performance in various domains. In this paper, we try to investigate the dimensionality reduction ability of auto-encoder, and see if it has some kind of good property that might accumulate when being stacked and thus contribute to the success of deep learning.

Based on the above idea, this paper starts from auto-encoder and focuses on its ability to reduce the dimensionality, trying to understand the difference between auto-encoder and state-of-the-art dimensionality reduction methods. Experiments are conducted both on the synthesized data for an intuitive understanding of the method, mainly on two and three-dimensional spaces for better visualization, and on some real datasets, including MNIST and Olivetti face datasets. The results show that auto-encoder can indeed learn something different from other methods. Besides, we preliminarily investigate the influence of the number of hidden layer nodes on the performance of auto-encoder and its possible relation with the intrinsic dimensionality of input data.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Artificial neural network is a mathematical and computational model composed of a large number of neurons that can simulate the structural and functional characteristics of biological neural network. It is a self-adaptive system, which changes the internal structure according to the external input, and is commonly used to model the complex relationship between input and output. The development of artificial neural network is a tortuous road. Perceptron is the starting point of modern neural computation, and the proof of perceptron convergence theorem in 1962 triggered the first climax of research for neural network represented by perceptron. In 1965, Minsky and Papert [2,20] pointed out the defects of perceptron and took a pessimistic view on the research of neural network, which made neural network study from rise into stagnation. By the early 1980s, related work by Hopfield et al. [16] showed the potential of neural network which made neural network study from stagnation to boom. In the mid-1990s, with the advent of the support vector machine (SVM) [10], researchers realized some limitations of artificial neural network and the research for neural network fell into low tide period again.

Auto-encoder, known as auto-association before, is a tricky three-layered neural network and was studied by a number of researchers in 1990s. Bourlard and Kamp [7] discussed the relationship between auto-association by multi-layer perceptrons and singular value decomposition in 1988. They showed that for auto-association with linear output units, the optimal weight values could be derived by purely linear techniques relying on singular value decomposition and low rank matrix approximation. In 1991, Kramer [18] introduced how to conduct nonlinear principal component analysis using auto-associative neural networks with three hidden layers. Due to the difficulty in training, deep network with multi-layer stacked auto-encoders did not exert its superior strength for a long time. Some researchers committed themselves to investigating several fundamental issues of neural networks with one or two hidden layers [8,1,22,23,3,9,17,25,21,27].

Until 2006, Geoffrey Hinton [15] solved this problem in Science Magazine to a large extent and broke the stalemate that neural network was in low tide. The method was to do layer-wise pre-training [14] to multi-layer auto-encoders using RBM [13]. That is the very hot topic in recent years—deep learning. From then on, deep learning sweeps across the industry and the academia like a wave. Recent research results have also demonstrated that deep learning indeed achieves state-of-the-art performances among various areas [4–6]. Particularly, deep learning has already been successfully applied to the industry in speech area. In the field of image analysis, there are also many good results. By building a

* Corresponding author.
E-mail addresses: wangyasi@hit.edu.cn (Y. Wang), h.yao@hit.edu.cn (H. Yao), zsc@hit.edu.cn (S. Zhao).

http://dx.doi.org/10.1016/j.neucom.2015.08.104
0925-2312/© 2015 Elsevier B.V. All rights reserved.

9-layered locally connected sparse auto-encoder, Le et al. [24] discovered that the network was sensitive to high-level concepts such as cat faces. Krizhevsky et al. [19] attained record-breaking performance on the ImageNet in 2012 with a large and deep convolutional neural network. Zeiler et al. [26] demonstrated state-of-the-art performance on Caltech-101 and Caltech-256 datasets. Zhu et al. [28] extracted face identity-preserving features from an image under any pose and illumination which could be used to reconstruct the face image in canonical view by designing a deep network.

Although achieving comparable performance and widely applied, deep learning is kind of like a "black-box" actually and there is no very sufficient and strict theoretical system to support. So a problem is: we have impressive performance using deep learning but we do not know why theoretically. In deep learning, a number of researchers tend to make progress by employing increasingly deep models and complex unsupervised learning algorithms.

This paper comes from the idea that whether auto-encoder has some kind of good property which might accumulate when being stacked and thus contribute to the success of deep learning. We start from a "building block" of deep learning—auto-encoder and focus on its dimensionality reduction ability. When restricting the number of hidden layer nodes less than the number of original input nodes in an auto-encoder, the desired dimensionality reduction effect can be achieved.

Based on the above analysis, the main contributions of the paper can be summarized as follows:

1. We start from auto-encoder and focus on its ability to reduce the dimensionality, trying to understand the difference between auto-encoder and state-of-the-art dimensionality reduction methods. The results show that auto-encoder indeed learn something different from other methods.
2. We preliminarily investigate the influence of the number of hidden layer nodes on the performance of auto-encoder on MNIST and Olivetti face datasets. The results reveal its possible relation with the intrinsic dimensionality.

On the whole, we expect that analyzing the fundamental methods in deep learning, e.g. auto-encoder, could help us understand deep learning better.

2. Auto-encoder based dimensionality reduction

In this section, we briefly introduce auto-encoder, four representative dimensionality reduction methods and the concept of dimensionality reduction and intrinsic dimensionality. Then we focus on auto-encoder's dimensionality reduction ability, and investigate the influence of the number of hidden layer nodes in auto-encoder.

2.1. Auto-encoder

Suppose the original input $\boldsymbol{x}$ belongs to n-dimensional space and the new representation $\boldsymbol{y}$ belongs to m-dimensional space, an auto-encoder is a special and tricky three-layered neural network in which we set the output $h_{W,b}(\boldsymbol{x}) = (\tilde{x_1}, \tilde{x_2}, \ldots, \tilde{x_n})^T$ equal to the input $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. J is the reconstruction error. It is an unsupervised learning algorithm and uses back propagation algorithm for training

$$h_{W,b}(\boldsymbol{x}) = g(f(\boldsymbol{x})) \approx \boldsymbol{x}$$

$$J(W, b; \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \| h_{W,b}(\boldsymbol{x}) - \boldsymbol{y} \|^2$$

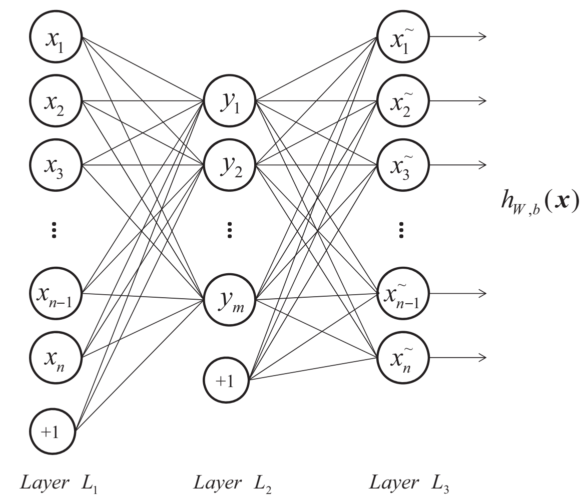

Fig. 1. The structure of auto-encoder.

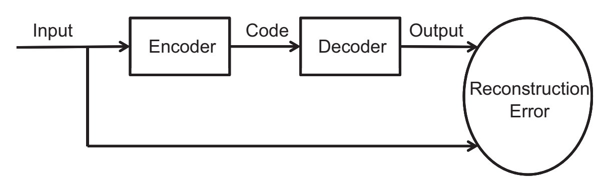

Fig. 2. The visualization description of auto-encoder.

Figs. 1 and 2 show the structure and the visualization description of an auto-encoder respectively. As shown in Fig. 2, from the first layer to the second layer amounts to an encoder f and from the second layer to the third layer amounts to a decoder g. We then minimize the reconstruction error J by adjusting parameters in the encoder and the decoder to get the code.

Auto-encoder can be seen as a way to transform representation. When restricting the number of hidden layer nodes m greater than the number of original input nodes n and adding sparsity constraint, the result will be similar to sparse coding. When restricting the number of hidden layer nodes m less than the number of original input nodes n, we can get a compressed representation of the input, which actually achieves desired dimensionality reduction effect. This paper mainly focuses on the latter case.

The following is a very quick introduction of four representative dimensionality reduction methods [12], which are to be used to compare with auto-encoder.

PCA: Principal component analysis is a very popular linear technique for dimensionality reduction. Given a dataset on $\mathbb{R}^n$, PCA aims to find a linear subspace of dimension d lower than n which attempts to maintain most of the variability of the data.

LDA: Linear discriminant analysis is another popular linear dimensionality reduction method. The basic idea is to ensure the samples after projection to have maximum-between-cluster-distance and minimum-in-cluster-distance in the new subspace.

LLE: Locally linear embedding is a nonlinear approach to reduce dimensionality by computing low-dimensional, neighborhood preserving embedding of high-dimensional data. A dataset of dimensionality n, which is assumed to lie on or near a smooth nonlinear manifold of dimensionality $d < n$, is mapped into a single global coordinate system of lower dimensionality d. The global nonlinear structure is recovered by locally linear fits.

Isomap: Isomap is a nonlinear generalization of classical multidimensional scaling. The main idea is to perform multidimensional scaling, not in the input space, but in the geodesic space of the nonlinear data manifold. The geodesic distance

Download English Version:

https://daneshyari.com/en/article/405956

Download Persian Version:

https://daneshyari.com/article/405956

Daneshyari.com